Segmentation of Dynamic Objects from Laser Data

Agustin Ortega and Juan Andrade-Cetto
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain

***Abstract*— We present a method to segment dynamic objects from high-resolution low-rate laser scans. Data points are tagged as static or dynamic based on the classification of pixel data from registered imagery. Per-pixel background classes are adapted online as Gaussian mixtures, and their matching 3D points are classified accordingly. Special attention is paid to the correct calibration and synchronization of the scanner with the the accessory camera. Results of the method are shown for a small indoor sequence with several people following arbitrarily different trajectories.**

***Index terms* – Segmentation, 3D sensing, calibration, sensor synchronization.**

I. Introduction

2D and 3D lidar scanning are popular sensing methodologies for robotics applications. They are used for robot navigation [5], trajectory planning [20], scene reconstruction [17], and even object recognition [1]. Aside from pricey devices such as the Velodyne HDL-64E, high resolution 3D lidar scanning is only possible at low frame rates. As an example, we have built an omnidirectional lidar sensing device for outdoor mobile robotics applications that scans with resolutions and acquisition times that range from 0.5 degrees at 9 seconds per revolution to finer point clouds sampled at 0.1 degrees resolution at a more demanding processing time of 45 seconds per revolution. This sensor has been devised for low cost, dense 3d mapping. The removal of dynamic and spurious data from the laser scan is a prerequisite to dense 3d mapping.

In this paper we address this problem by synchronizing the laser range sensor with a color camera, and using the high frame-rate image data to segment out dynamic objects from the point clouds. Per-pixel class properties of image data are adapted online using Gaussian mixtures. The result is a synchronized labeling of foreground/background corresponding laser points and image data as shown in Fig. 1.

The paper is organized as follows. In the next section we present related work in background segmentation using computer vision methods and 3D laser range data. Section III gives our custom built sensor specifications, and details the methods developed for sensor synchronization and sensor calibration. Section IV details the background segmentation algorithm. Results of the method are shown in Section V on a real indoor scenario with several people moving with random patterns. Conclusions and future work are detailed in Section VI.

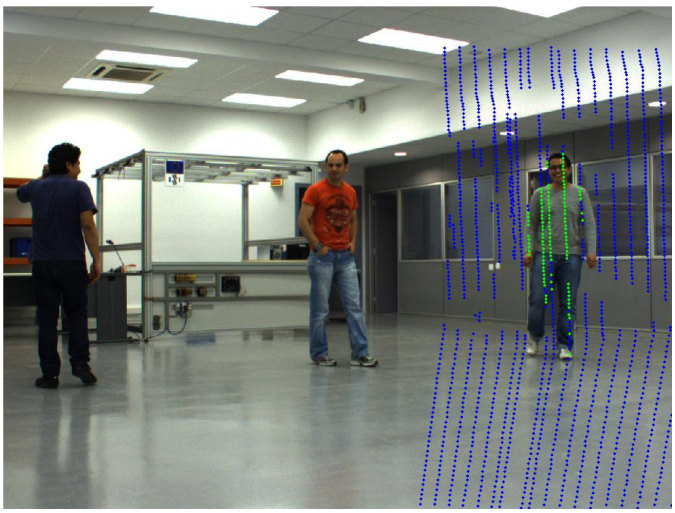


Fig. 1. Several laser scans of a dynamic object reprojected on their corresponding image frame.

II. Related Work

Methods that study the segmentation of 3D laser data usually focus on the extraction of valuable geometric primitives such as planes or cylinders [10] with applications that vary from map building, to object classification [2], road classification [6], or camera network calibration [9]. All these methods however are designed to work on static data only and do not consider the temporal information. For outdoor map building applications, the removal of dynamic objects from the laser data is desirable. Furthermore, for low-rate scanning devices such as ours, moving items in the scene would appear as spurious 3D data; hence the need to segment them out.

Background segmentation is a mature topic in computer vision, and is applied specially to track objects in scenarios that change illumination over time but keep the camera fixed to a given reference frame. The most popular methods adapt the probability of each image pixel to be of background class using the variation of intensity values over time. Such adaptation can be tracked with the aid of a Kalman filter [14] taking into account illumination changes and cast shadows. These methods can be extended to use multimodal density functions [18, 19] in the form of Gaussian mixture models, whose parameters are updated depending on the membership degree to the background class.

The classification of 3D range data fusing appearance information has been addressed in the past, again for static scene analysis. Posner et al. [11, 12, 13] propose an unsupervised

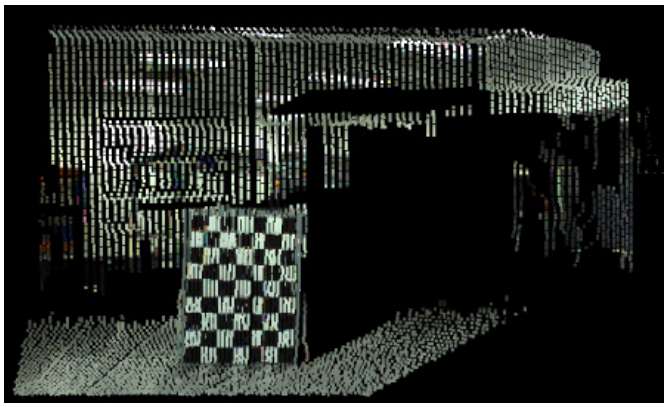


Fig. 2. Camera to laser rigid body pose estimation using a planar calibration pattern.

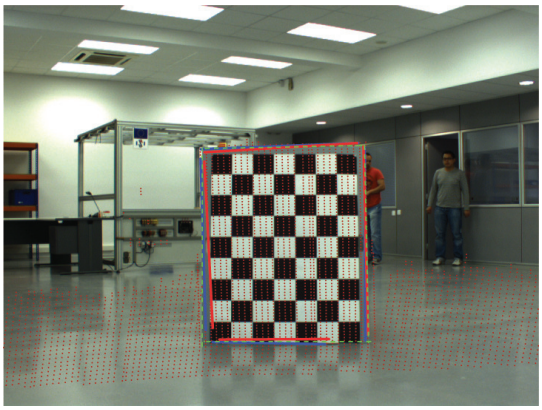


Fig. 4. Laser-camera pose refinement using line primitives. The green dotted lines show the image features. Red lines show reprojection prior to pose refinement, and blue lines correspond to refined reprojected estimates (best viewed in color).

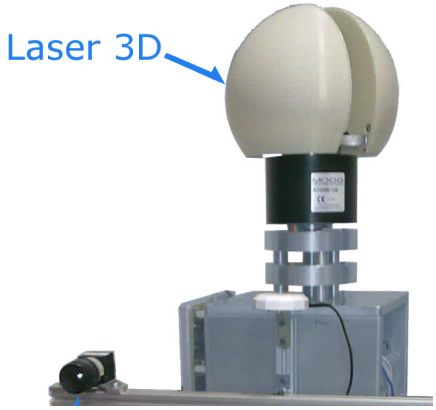


Fig. 3. Our custom built 3D range sensing device and a rigidly attached color camera.

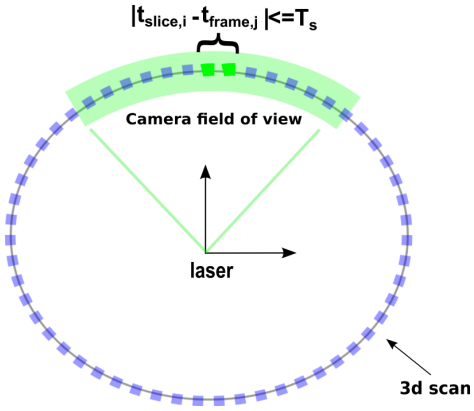


Fig. 5. Camera and laser synchronization.

method that combines 3D laser data and monocular images to classify image patches to belong to a set of 8 different object classes. The technique oversegments images based on texture and appearance properties, and assigns geometric attributes to these patches using the reprojected 3D point correspondences. Each patch is then described by a bag of words and classified using a Markov random field to model the expected relationship between patch labels.

These methods (and ours) have as a prerequisite the accurate calibration of both sensors, the laser and the camera. The computation of the rigid body transformation between 2D and 3D laser scanners and a camera are common procedures in mobile robotics and are usually solved with the aid of a calibration pattern. The techniques vary depending on the type of sensor to calibrate, and on the geometric motion constraints between the two sensor reference frames [23, 22, 7, 9]. Sensor synchronization on the other hand has received less attention. Sensor synchronization and occlusions are studied in [16] for the case of the Velodyne HDL-64 sensor. A more general method to synchronize sensors with varying latency is proposed in [8].

III. Sensor Synchronization and Calibration

A. Sensor specifications and data acquisition

Our 3D range sensing device consists of a Hokuyo UTM-30LX laser mounted on a slip-ring, with computer-controlled angular position via a DC brushless motor and a controller. For the experiments reported in this paper, laser resolution has been set to 0.5 degrees in azimuth with 360 degree omnidirectional field of view, and 0.5 degrees resolution in elevation for a range of 270 degrees. Each point cloud contains 194,580 range measurements of up to 30 meters with noises varying from 30mm for distances closer to 10m, and up to 50mm for objects as far as 30m. The color camera used is a Pointgray Flea camera with M1214-MP optics and 40.4 degree field of view. Fig. 3 shows a picture of the entire unit.

B. Sensor Calibration

We are interested in the accurate registration of laser range data with intensity images. Registration can be possible by first calibrating the intrinsic camera parameters and then, finding the relative transformation between the camera and laser reference frames. Intrinsic camera calibration is computed using Zhang's method and a planar calibration pattern [24],

although other methods could also be used [3, 21]. Extrinsic calibration between the laser and camera is initialized by selecting correspondences of the calibration plane corners on both sensing modalities with the aid of a graphical user interface, and using Hagger's method for pose estimation [4], as shown in Fig. 2.

The method is subject to the resolution of the laser scanner for the selection of the four 3D to 2D corner matches in the pattern. Pose estimation is further refined by minimizing the reprojection error of line primitives. Lines in the 3D point cloud are obtained growing and intersecting planar patches as in [10]. Their corresponding matches in the images are manually selected using the graphical user interface.

Line reprojection error is computed as the weighted sum of angular and midpoint location reprojection errors as shown in Fig. 4,

$$\epsilon = \sum (\theta_i - \theta_p)^2 + w(m_i - m_p)^T (m_i - m_p) \qquad (1)$$

where the subscript i corresponds to measured image features, and the subscript p indicates projected model features. The weight w is a free tuning parameter to account for the difference between angular and Cartesian coordinates.

C. Synchronization

At 0.5 degree resolution, our 3D scanner takes about 9 seconds to complete a scan, which is made of a 180 degree turn of the sensor. Camera frame rate is set to 17 fps, thus we have roughly 153 images per full 3D image.

The timestamps between consecutive laser slices $t_{\texttt{slice}_i}$, and grabbed images $t_{\texttt{frame}_j}$ are compared and set to lie within a reasonable threshold T_s in milliseconds.

$$|t_{\texttt{slice}_i} - t_{\texttt{frame}_j}| \leq T_s \qquad (2)$$

With $T_s = 1/17$, each laser scan is uniquely assigned to its corresponding image frame, roughly two to three per image. Increasing this threshold, allows to match each laser slice to more than one image at a time (see Fig. 5).

IV. Background Sustraction

Once we have time correspondences between 3D laser slices and image frames, we can use background segmentation results on the image sequence to classify the corresponding 3D points in each time slice as belonging to a dynamic or static object. The method we implemented is based in [19].

A. Mixture Model

For each pixel in the image, the probability of its RGB coordinates x to be of the background class is modeled as a mixture of K Gaussian distributions.

$$p(x) = \sum_{k=0}^{K} \omega_k \mathcal{N}(x|\mu_k, \Sigma_k) \qquad (3)$$

with ω_k the weight of the k-th Gaussian, and K a user selected number of distributions.

This classification scheme assumes that the RGB values for neighboring pixels are independent. During the training session, when a pixel RGB value x falls within 2.5 standard deviations of any of the distributions in the sum (in the Mahalanobis sense), evidence in the matching distributions is stored by recursively updating their sample weight, mean, and variance with

$$\omega_k(t+1) = (1-\alpha)\omega_k(t) + \alpha \qquad (4)$$

$$\mu_k(t+1) = (1-\rho)\mu_k(t) + \rho x \qquad (5)$$

$$\Sigma_k(t+1) = (1-\rho)\Sigma_k(t) + \rho(x-\mu(t))^T(x-\mu(t)) \qquad (6)$$

and

$$\rho = \alpha \mathcal{N}(x|\mu_k, \Sigma_k) \qquad (7)$$

Note that after updating ω in Eq. 4, the weights need to be renormalized. And, just as in [19] we also consider during the training session, that when a pixel value x falls below a 2.5 standard deviation of the distribution, the least probable distribution of the Gaussian sum is replaced by the current RGB pixel value as the current mean, with an initially high variance, and a low prior weight.

B. Background Class

The mixture model on each pixel encodes the distribution of colors for the full image sequence set per full 3D scan (about 153 images). The static portion of the data, i.e., the background, is expected to have large frequency and low variance. By ordering the Gaussians on each sum by the value $\frac{\omega}{\det \Sigma}$, the distributions with larger probability to be of the background class will be aggregated in the top of the list. Static items might however be multimodal in their color. For instance, a flickering screen or a blinking light. As a result we choose as background class the first $B < K$ ordered distributions which add up to a factored weight ω_B, where

$$B = \texttt{argmin}_b \left(\sum_{i=1}^{b} \omega_i \geq \omega_B\right) . \qquad (8)$$

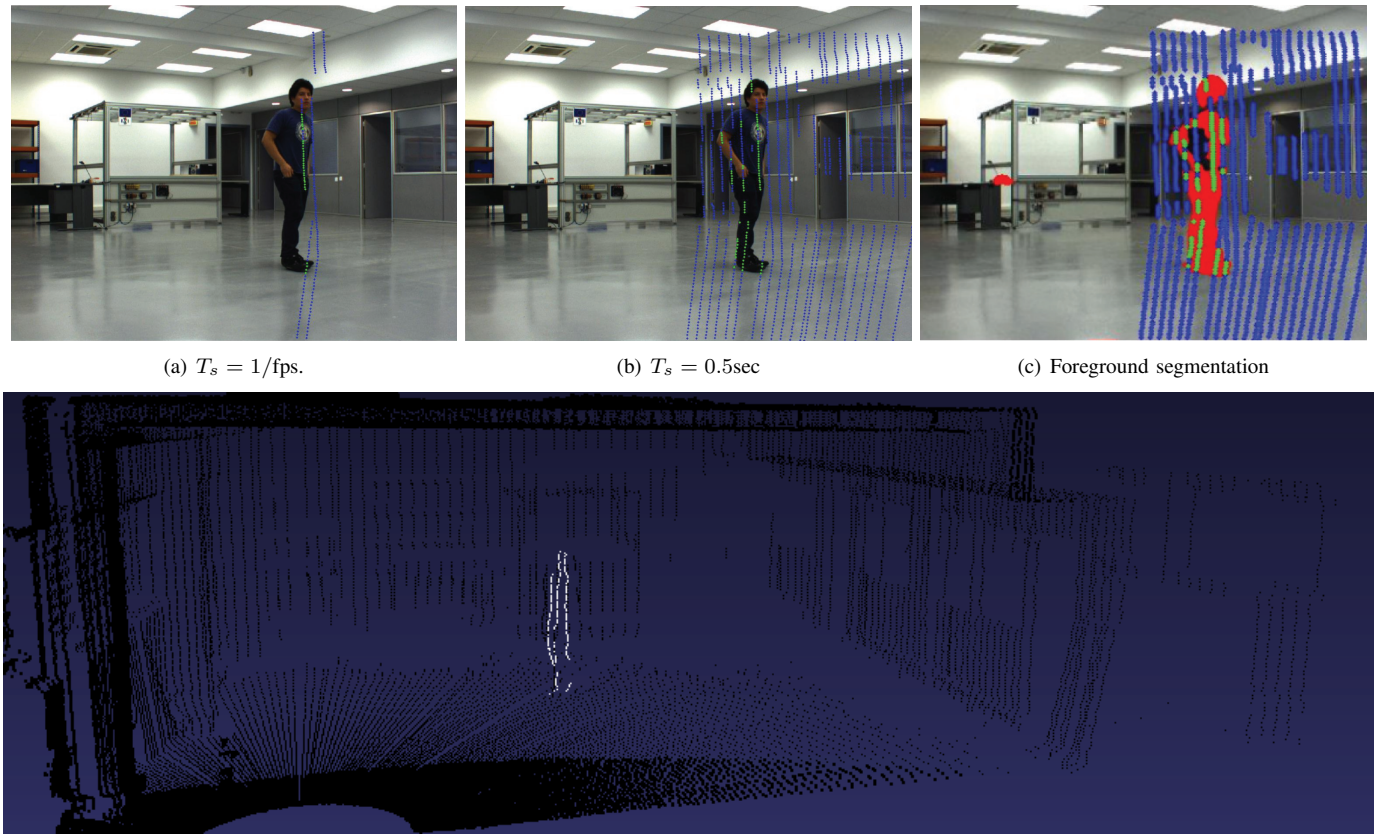
C. Point classification

Each point on each scan slice is reprojected to its matching image frames according to Eq. 2. Ideally, for tight bounds on T_s, only one image will be assigned to each scan slice. Robustness to noise is possible however, if this bound is relaxed and we allow for larger values of T_s, so that more than one image can be matched to the same scan slice. We call this set of images I.

Thus for each point in a slice, the corresponding pixel values x from the whole set I is visited, and checked for inclusion in the set B of distributions. Class assignment is made if x belongs to B for all the images in the set I.

V. Experiments

Results are shown for a series of indoor sequences with moderate dynamic content. For background segmentation, the multimodal distribution is set to contain 4 Guassians, the learning rate is set at $\alpha = 0.3$, and the background class is set to one third of the frequency in the distributions, i.e.,



(a) $T_s = 1/\text{fps}$. (b) $T_s = 0.5\text{sec}$ (c) Foreground segmentation

(d) Segmented dynamic object

Fig. 6. Segmentation results for a sequence with one moving person and varying values of the synchronization threshold.

$\omega_B = 0.3$. The synchronization threshold T_s is varied from the minimal $1/17$ to a more conservative value of 0.5 seconds.

The first analyzed sequence corresponds to a single person moving in front of the laser and camera. Frames (a) and (b) in Figure 6 show final results of point classification for different values of T_s; frame (c) shows the image pixel classification results; and frame (d) shows the 3D reconstruction of both, the segmented dynamic object, and the entire 3D scene.

The second sequence contains a more challenging scenario with three people with slow random walking trajectories. Given the slow motion rate of the people, laser range readings hitting on them are difficult to categorize as being dynamic. The background segmentation algorithm proposed in this paper helps to alleviate this issue. Figure 7 shows results of background segmentation in this new sequence for varying values of the synchronization parameter. Setting this parameter slightly above the camera acquisition rate accounts for synchronization errors and produces better segmentation results. Frames (a-c) in the image show the segmentation results for $T_s = 1/\text{fps}$, whereas frames (d-f) show segmentation results for $T_s = 0.5\text{sec}$.

Figure 8 shows 3D reconstruction results of the segmented data and of the full 3D scene. The results shown are for a synchronization threshold of 0.5 sec.

We appreciate the suggestion during the peer review phase of this work to compare our method with other approaches. Unfortunately, as far as we know, the system presented is unique, and there are no other methods in the literature that take low-rate 3d scans and remove dynamic content from them using high-rate imagery. To validate the approach, we can report however an empirical comparison with ground truth image difference. Assuming a clean background scan is available (without people), image difference to a full dynamic cloud was computed with the Point Cloud Library [15] using a distance threshold of 3mm. Fig. 9 shows results of such image difference computation. The results of our method are visually comparable to such ground truth experiment.

VI. Conclusions

We present a method to segment low-rate 3D range data as static or dynamic using multimodal classification. The technique classifies fast-rate image data from an accessory camera as background/foreground adapting at frame rate a per-pixel Gaussian mixture distribution. The results of image classification are used to tag reprojected laser data.

Special attention is paid to the synchronization and metric calibration of the two sensing devices. Sensor synchronization is of paramount importance as it allows to match high frame rate imagery with their corresponding low rate laser scans. The method is tested for indoor sequences with moderate dynamics.

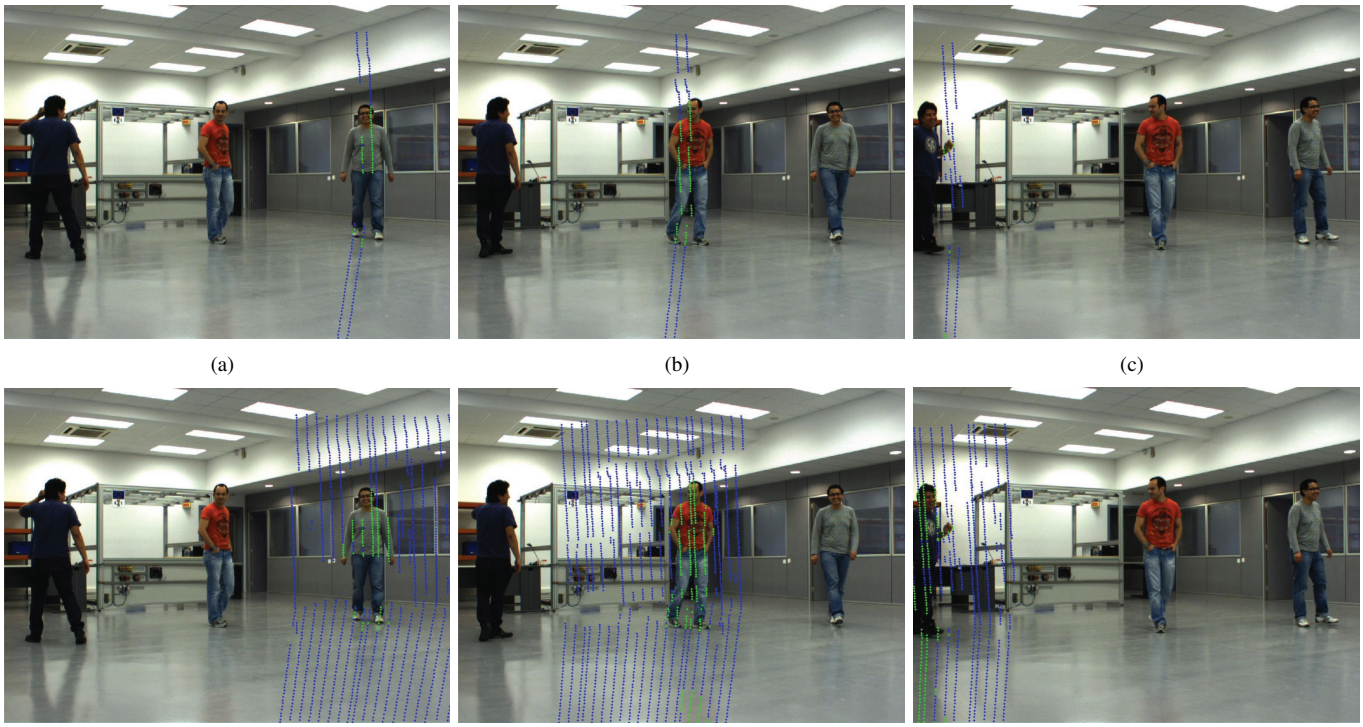


Fig. 7. Segmentation results for a sequence with three people moving randomly and varying values of the synchronization threshold. Frames (a-c) show three sequence instances segmented at $T_s = 1/\text{fps}$. Frames (d-f) show the same sequence instances segmented at $T_s = 0.5\text{sec}$.

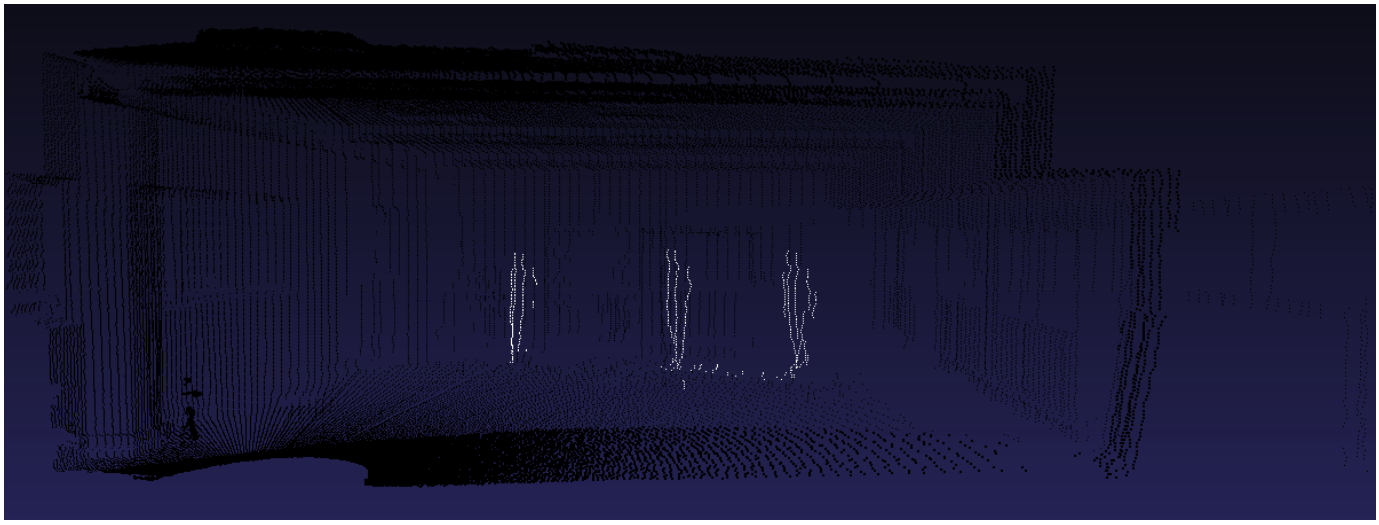


Fig. 8. Segmentation results for a sequence with three slowly moving people with random walking trajectories.

The proposed method was designed to remove spurious data or dynamic objects from low acquisition rate lidar sensors. The result is a cleaner 3d picture of static data points. These point clouds could then be aggregated into larger datasets with the guarantee that dynamic data and noise will not jeopardize point cloud registration. The intended application of the technique is robotic 3d mapping.

VII. Acknowledgments

This work has been partially supported by the Mexican Council of Science and Technology with a PhD Scholarship to A. Ortega, by the Spanish Ministry of Science and Innovation under projects PAU (DPI2008-06022) and MIPRCV Consolider Ingenio (CSD2007-018), and by the CEEDS (FP7-ICT-2009-5-95682) and INTELLACT (FP7-ICT2009-6-269959) projects of the EU. The authors thank M. Morta for the development of the 3D laser used for this research.

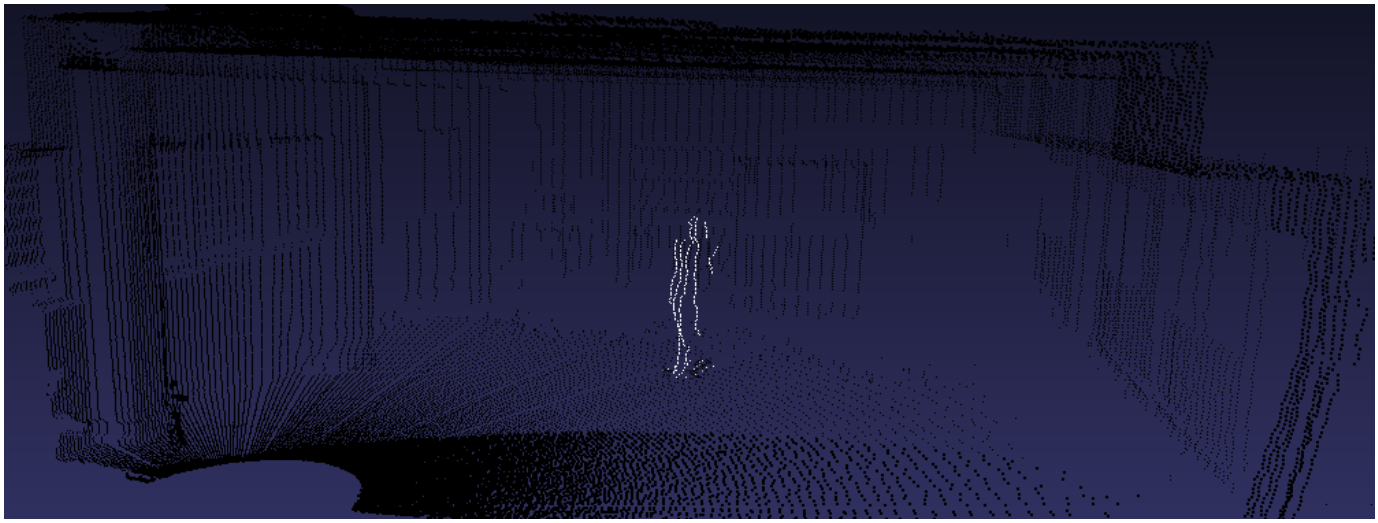


Fig. 9. Result of applying point cloud difference using PCL.

REFERENCES

[1] K.O. Arras, O.M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3402–3407, Rome, April 2007.

[2] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard. Unsupervised discovery of object classes from range data using latent Dirichlet allocation. In *Robotics: Science and Systems V*, Seattle, USA, June 2009.

[3] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2nd edition, 2004.

[4] C.P. Lu, G.D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:610622, 2000.

[5] F. Maurelli, D. Droeschel, T. Wisspeintner, S. May, and H. Surmann. A 3D laser scanner system for autonomous vehicle navigation. In *Proceedings of the 14th International Conference on Advanced Robotics*, Munich, June 2009.

[6] F. Moosmann, O. Pink, and C. Stiller. Segmentation of 3D lidar data in non-flat urban environments using a local convexity criterion. In *IEEE Intelligent Vehicles Symposium*, pages 215–220, 2009.

[7] P. Nú nez, P. Drews Jr, R. Rocha, and J. Dias. Data fusion calibration for a 3D laser range finder and a camera using inertial data. In *Proceedings of the European Conference on Mobile Robotics*, Dubrovnik, September 2009.

[8] E. Olson. A passive solution to the sensor synchronization problem. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1059–1064, Taipei, October 2010.

[9] A. Ortega, B. Dias, E. Teniente, A. Bernardino, J. Gaspar, and Juan Andrade-Cetto. Calibrating an outdoor distributed camera network using laser range finder data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 303–308, Saint Louis, October 2009.

[10] A. Ortega, I. Haddad, and J. Andrade-Cetto. Graph-based segmentation of range data with applications to 3d urban mapping. In *Proceedings of the European Conference on Mobile Robotics*, pages 193–198, Dubrovnik, September 2009.

[11] I. Posner, M. Cummins, and P. Newman. Fast probabilistic labeling of city maps. In *Robotics: Science and Systems IV*, Zurich, June 2008.

[12] I. Posner, D. Schroeter, and P. Newman. Describing composite urban workspaces. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4962–4968, Rome, April 2007.

[13] I. Posner, D. Schroeter, and P. Newman. Online generation of scene descriptions in urban environments. *Robotics and Autonomous Systems*, 56(11):901–914, 2008.

[14] C. Ridder, O. Munkelt, and H. Kirchner. Adaptive background estimation and foreground detection using kalman-filtering. In *Proceedings of the IASTED International Conference on Robotics and Manufacturing*, pages 193–199, Istanbul, August 1995.

[15] R.B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *Proceedings of the IEEE International Conference on Robotics and Automation*, Shanghai, May 2011.

[16] S. Schneider, M. Himmelsbach, T. Luettel, and H.J. Wuensche. Fusing vision and lidar - synchronization, correction and occlusion reasoning. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 388–393, San Diego, June 2010.

[17] I. Stamos and P.K. Allen. 3-d model construction using range and image data. In *Proceedings of the 14th IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 1531, Hilton Head, SC, June 2000.

[18] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the 13th IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, Fort Collins, June 1999.

[19] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[20] H. Surmann, A. Nuchter, and J. Hertzberg. An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments. *Robotics and Autonomous Systems*, 45(3-4):181–198, 2003.

[21] R. Tsai. A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras. *IEEE Journal of Robotics and Automation*, 3(4):323–344, August 1987.

[22] R. Unnikrishnan and M. Hebert. Fast extrinsic calibration of a laser rangefinder to a camera. Technical Report CMU-RI-TR-05-09, Robotics Institute, Pittsburgh, July 2005.

[23] Q. Zhang and R. Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2301–2306, Sendei, September 2004.

[24] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.